IT LOAD

The power requirements of servers depend on the actual computing load: Therefore the usage patterns of the equipment that describe the IT load processed are needed. In real Data Centres, disparities in performance and power characteristics across servers and different scheduling, task migration or load balancing mechanisms, have effects on its power consumption that are difficult to predict. For simplicity, three different homogeneous IT load (Web, HPC and Data) have been studied in the present manuscript:

- **Web workload** has real-time requirements: the users of such workload need to get a response to their petitions in few seconds (i.e. Google search, Facebook surf, etc.). There is not a typical resource consumption profile for web workloads but they may use CPU, memory, network or disk in several proportions. This workload has the particularity to follow a daily/weekly pattern.
- **HPC workload** is typically CPU intensive. They perform a large amount of floating-point operations for scientific calculations. Because HPC workloads may last for hours, or even days, they do not have real-time requirements, and they are usually allocated in job queues that may execute them hours or days after they are submitted by the users.
- **Data workload** is usually both memory and disk-intensive, while they can also use a high rate of CPU operations for data analysis. Despite of data workloads may have real-time requirements (i.e. a search query in Google), the authors consider data workloads without real-time requirements (i.e. background data analytics for business intelligence applications).

Notice that HPC and Data workloads do not follow a given pattern, and they will depend on the access policy and dimension of the data centres of each institution. In the framework of the project, Web workload is a real pattern collected from the access log of an ISP within the UPC [1], while HPC and data workloads patterns are extracted from the CEA-Curie data centre which are publicly available in the Parallel Workloads Archive [2]. Figure 1 and 2 presents the three IT load profiles over a week.

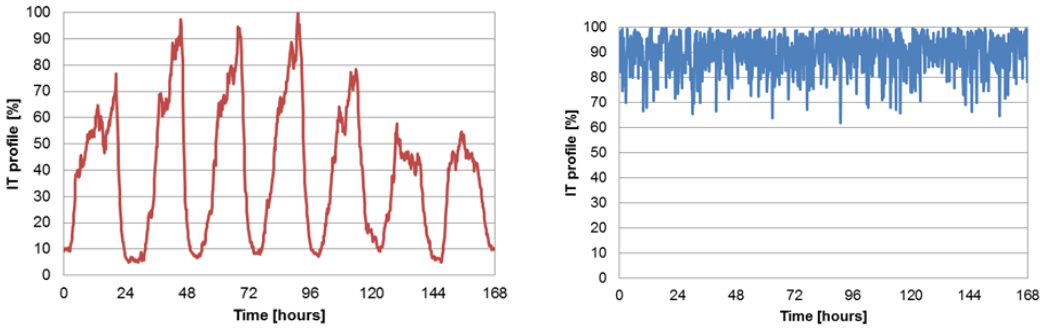

Figure 1 IT workload profiles during a week. Left: Web profile, Right: Data profile.

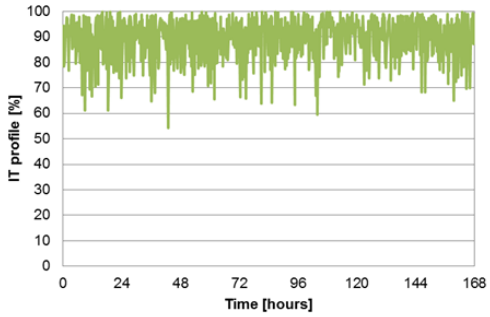

Figure 2 HPC workload profile during a week.

However, in order to predict the data centre consumption from the IT load, a relationship between server usage (IT load) and server consumption (IT power) is developed. The definition of IT load is an additive function that considers the load rates of CPU, Main Memory, Disk and Network, pondered according to the measured impact of each term in late 2012 servers [3]. Firstly, different types of micro-benchmark for fully stressing the system were executed in order the reach the maximum real power of the system. These benchmarks included Ibench suite [4], Stress-ng [5], Sysbench [6], Prime95 [7], Pmbw [8], Fio [9] and Iperf3 [10]. After this initial process different benchmarks based on real-world software stacks from CloudSuite [11] for web and data benchmarks, and NAS Parallel Benchmarks [12] for HPC were also executed. With this experimentation the relation between IT load and power consumption has been derived. Notice that for its further adaptation to other hardware these correlations were normalized. Figure 3 shows the results of the experimentation and the regressions to predict different consumptions in function of the IT load. The variability in the power/load measurements show that there is not a generic power profile for software, because all the components of a host (CPU, memory, disk, network) do not work independently. They must coordinate because there are dependencies between data and procedures (and the usage of resources

is variable across the same execution, depending of the availability of their required inputs at a given time).

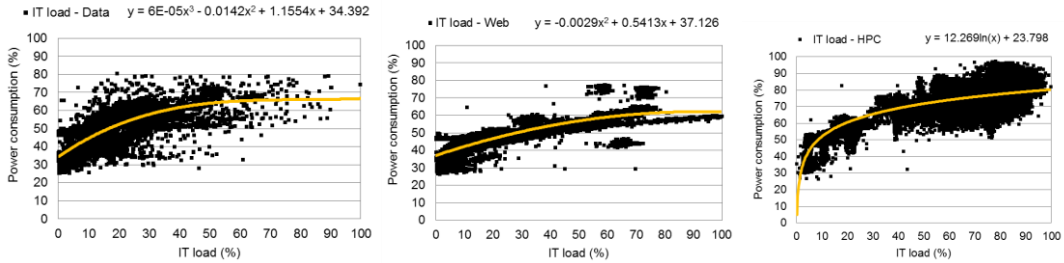

Figure 3 Curves for the three types of workloads and their corresponding regression. The y-axis shows the percentage of the power referred to the maximum detected. The x-axis shows the percentage of the IT load referred to the maximum executed.

In the project, 3 different IT load scenarios are available:

- Data centre dedicated totally to **HPC workload**
- Data centre dedicated totally to **Web workload**.
- **Mixed used data centre**; in this scenario IT load is composed by 40 % web, 30 % HPC and 30 % data workload.

Figure 4 presents the servers consumption in percentage of the total installed IT capacity for the three scenarios presented. Notice that even though HPC workload is working nearly at 100% of load, this not represents 100% of the IT equipment consumption (at maximum 80%). The maximum values of % IT power consumption reached for Web and Mixed uses is lower than 70%.

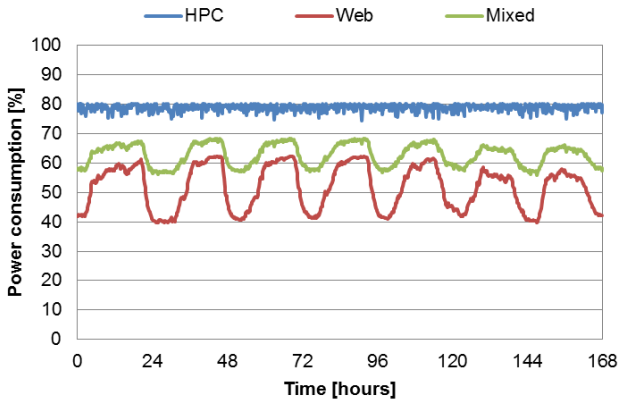

Figure 4 Percentage of the maximum IT power consumption for different IT load distribution cases.

REFERENCES

[1] M. Macías and J. Guitart, "SLA negotiation and enforcement policies for revenue maximization and client classification in cloud providers," *Future Gener. Comp. Sy.,* vol. 41, p. 19–31, 2014.

[2] [Online]. Available: http://www.cs.huji.ac.il/labs/parallel/workload/.

[3] [Online]. Available: http://www.morganclaypool.com/doi/pdfplus/10.2200/S00516ED2V01Y201306CAC024.

[4] C. Delimitrou and C. Kozyrakis, "iBench: Quantifying interference for datacenter applications," in *2013 IEEE International Symposium on Workload Characterization (IISWC)*, 2013.

[5] [Online]. Available: http://kernel.ubuntu.com/~cking/stress-ng/.

[6] [Online]. Available: https://launchpad.net/sysbench.

[7] [Online]. Available: http://www.mersenne.org/download/.

[8] "Pmbw - Parallel Memory Bandwidth benchmark / measurement," [Online]. Available: http://panthema.net/2013/pmbw/.

[9] "Fio - Flexible I/O tester," [Online]. Available: http://git.kernel.dk/?p=fio.git;a=summary.

[10] "Iperf3: A TCP, UDP, and SCTP network bandwidth measurement tool," [Online]. Available: https://github.com/esnet/iperf.

[11] [Online]. Available: http://parsa.epfl.ch/cloudsuite/cloudsuite.html.

[12] [Online]. Available: http://www.nas.nasa.gov/publications/npb.html.